EEGym: a novel benchmark for comparing language model representations and human N400 responses

Sam Hutchinson [1] **Jenny Baek** [1] **Ali Cy** [1]

Abstract

Despite the success of LLMs at generating human-like text, the extent of the similarity between their internal representations and those in the human mind and brain remains a matter of debate. One way researchers have approached this question is to test the extent to which the representations from different models can predict human neural activity. One classic neural signature of human language processing is the N400 ERP, and electrical signal measured with EEG, and there have been several competing claims in the literature in recent years regarding how best to computationally model the N400 using LLMs. These experiments, however, largely use different datasets, experimental paradigms, and comparison models. To shed some light on this debate, we compile EEGym: a standardized ERP dataset with data from three separate lab experiments, totaling nearly 46,000 individual trials. We then compare representations from a suite of models in their ability to predict these responses. We replicate previous results that find LLM *surprisal* to be the best predictor of the N400, and show the novel result that this predictivity scales with model size. We also find that, at least for the Qwen family of models, pre-trained-only models are better predictors of human neural activity than instruction-tuned and post-trained models. Cleaned datasets, analysis code, and interactive figures can be found at: https://samhutch511.github.io/nlp-project/.

1. Introduction

In recent years, LLMs have shown the ability to produce convincing human language, prompting many researchers to compare the internal representations of LLMs to human neural representations during language comprehension (Schrimpf et al., 2021; Caucheteux et al., 2023; Tuckute et al., 2024). This endeavor is two-fold. One, LLM models could be better aligned with human-relevant tasks if their model representations of language are more "human-like." Two, such models could help broaden our understanding of human language processing. Psycholinguists have investigated language processing at the neural level by studying event-related-potentials (ERPs), which are instances of brain activity that are directly triggered by cognitive, sensory, or motor events. The N400 ERP, which is thought to be triggered by semantic processing (Kutas & Federmeier, 2011), is one of the most reliable and well-studied measures of language comprehension. N400 has been reported to be sensitive to various behavioral metrics of word-expectancy, such as cloze probability, plausibility ratings, or reading times. This holds across both naturalistic sentence reading tasks and word priming paradigms (Kutas & Hillyard, 1984; Rugg, 1985; Michaelov et al., 2024). With LLMs, one metric of word-expectancy is LLM surprisal, which is the predictability of a word in a given context. A specific example is last-layer LLM surprisal, which measures the predictability of a token from the probability distribution over tokens established by the very last LLM layer. More recently, several studies have found that LLM surprisal is correlated with N400 amplitudes as well (Frank et al., 2015; Michaelov et al., 2024).

However, simply associating last-layer LLM surprisal with the N400 response poses several theoretical questions left open in the literature (de Varda et al., 2024). For instance, LLM surprisal is strictly a measure of the prediction error over the next token (or word), whereas the N400 has been shown to be more broadly sensitive to the semantic content of non-linguistic symbols or events (Kutas & Federmeier, 2011). Additionally, LLM surprisal conflates two features which have both been shown to drive N400 responses: the expectancy (or contextual plausibility) of a word, and that word's semantic association with the preceding context. Sentences with words which are contextually implausible but semantically associated have been shown to elicit smaller N400s than sentences with words which are semantically unrelated and implausible (Krieger et al., 2025; Aurnham-

[1]6.8610 Grad NLP, MIT. Correspondence to: Sam Hutchinson <samhutch@mit.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

Table 1. Datasets included in EEGym, with example items

DATASET	TYPE	N. SUBJECTS	N. ITEMS	EXAMPLE ITEMS ("ITEM" / ***critical word***)
FRANK-2015	NATURALISTIC	24	1668	"A HORSE HAS THROWN A SHOE." / ***horse***
				"A HORSE HAS THROWN A SHOE." / ***has***
RYSKIN-2021	CONTROLLED	24	640	"AFTER SHE DECEIVED HER MOTHER, SHE HAD A GUILTY MUSKET." / ***musket***
				"POLITICAL CORRUPTION AND ECONOMIC HARDSHIP FUELED A GROWING ANTI-GOVERNMENT DALMATIAN." / ***dalmatian***
MICHAELOV-2024	CONTROLLED	50	500	"IF JIM'S MANAGER IS INCOMPETENT, HE HAS A LEGITIMATE RIGHT TO WATER." / ***water***
				"NORTON HATES GETTING UP ON WINTER MORNINGS BECAUSE IT IS ALWAYS BROKEN." / ***broken***

mer et al., 2021). To address these theoretical issues, several recent studies have proposed alternative, competing metrics to LLM surprisal in order to algorithmically explain N400 responses (de Varda et al., 2024, Krieger et al., 2025; Li & Futrell, 2024; Lopopolo & Rabovsky, 2024; Eddine et al., 2022). However, these studies compare models against different datasets and experimental designs, making direct comparison of their claims difficult.

We propose EEGym as an initial foray into the creation of a benchmark for comparing language model representations and human N400 responses. We create this benchmark by aggregating and standardizing data across several experimental datasets and comparing representations from a consistent set of models. This will allow us to test the various predictions extant in the literature referenced above: for example, whether representations from Transformers or LSTMs will be better predictors (i.e. Michaelov et al., 2024 versus Lopopolo & Rabovsky, 2024), and whether the standard next-token surprisal metrics or more neuro-cognitively-inspired metrics, such as representations from units selected with a "localizer" framework, will be better predictors (i.e. Li & Futrell, 2024 and AlKhamissi et al., 2025). Varying the model size and training types within model families will also allow us to test the impact of model scale on brain predictivity (Antonello et al., 2023). For the theoretical reasons outlined above, we expect the cognitively-inspired metrics from larger transformer models will outperform both raw last-layer surprisal and LSTM representations in terms of predicting N400 responses.

2. Related Work

Standard NLP benchmarks (GLUE, BLiMP, SuperGLUE) measure whether models produce correct outputs on held-out tasks; reading-time corpora (PROVO, Dundee, MECO) capture a coarse measure of behavioral speed but do not capture more specific computations. The N400, however, is an index of real-time word-by-word semantic processing, allowing for an assessment of whether models are able to process the meaning of language in a human-like way. One issue with comparing competing claims made by the various papers cited above is that they use divergent datasets. A standardized N400-based NLP benchmark will allow us to more rigorously test cognitive claims about LLMs, as well as their comparability to human language processing. Finally, our expanded set of comparison metrics and representations-under-investigation is a novel contribution to the field of cognitively-inspired human-LLM comparisons, and has the potential to advance our understanding of the computations and representations underlying human language processing.

3. Methods

3.1. Dataset Aggregation

In order to rigorously compare model representations and human N400 responses in a standardized way, we need datasets which meet several criteria: 1) the data must be accessible from an open-source repository, such as OSF or GitHub; 2) the data must be reported trial-wise, with the N400 amplitude reported for each participant on each trial; 3) the datasets must report the entire sentence for each trial, as well as the critical word; and 4) the dataset must include indications of which trials to exclude based on artifacts or other recording errors. Intersecting this list of criteria with the datasets used by the competing papers in the recent literature leaves us with three experimental datasets: from Frank et al. 2015; Ryskin et al. 2021; and Michaelov et al. 2024.

In order to wrangle these datasets into a standardized format, we first removed all trials marked as exclusions in the base datasets, we then averaged each subject's responses in each trial over the amplitude recorded at the standard set of centro-parietal electrodes over which the N400 is typically defined (Kutas & Federmeier, 2011), and then finally

Z-scored the responses within each base experiment. This final step is necessary to account for variations between labs and recording equipment, which can affect the absolute magnitude of responses.

This cleaning process yields a dataset of N400 responses to individual words with data from 98 subjects and 45,879 individual trials. The data is summarized, with example items, in Table 1. Frank et al. 2015 collected naturalistic reading data from participants continuously narrative texts and reported the EEG amplitude at each word in each sentence, as shown in Table 1. Ryskin et al. 2021 and Michaelov et al. 2024, on the other hand, collected controlled datasets with participants reading one sentence at a time, across several different violation conditions. Table 1 shows two examples of *semantic* violations from each dataset. Both datasets report the EEG amplitude at the critical anomalous word, which is always the final word in the sentence. We lowercase the entire sentence and tokenize the lowercased sentence for presentation to the models, as discussed below. If a critical word spans multiple tokens, we sum their surprisals, as in de Varda et al. 2024.

3.2. Model Selection

In addition to comparing across the same data, a fair comparison of the existing proposals for computational accounts of the N400 should use the same models. We therefore selected a set of models to use across all of the datasets, based on several criteria: 1) open-source model access; 2) some standard for use in cognitive neuroscience, NLP, or interpretability research; 3) varied model sizes within the same model family, in order to investigate the impact of scale; and 4) varied training types within the same model family and size, in order to investigate the impact of pre-training versus post-training. We list the models we chose and briefly describe their similarities and differences below:

- **GPT-2** (Radford et al., 2019): A suite of decoder-only autoregressive transformers introduced by OpenAI and trained on WebText, a 40 GB corpus of text scraped from outbound links on Reddit. We use three sizes—GPT-2 (117M parameters, 12 layers), GPT-2-Large (774M, 36 layers), and GPT-2-XL (1.5B, 48 layers). GPT-2 is often used as a standard comparison model in psycholinguistic and cognitive neuroscience research, particularly in fMRI encoding models (e.g. Tuckute et al., 2024; Schrimpf et al., 2021).

- **Pythia** (Biderman et al., 2023): A suite of decoder-only autoregressive transformers developed by EleutherAI and trained on the deduplicated Pile, a 825 GB diverse English text corpus. We use five sizes—410M (24 layers), 1.4B (24 layers), 2.8B (32 layers), 6.9B (32 layers), and 12B (36 layers). Pythia was designed explicitly for mechanistic interpretability: all model sizes share the same training data, code, and architecture family, differing only in depth and width, which helps us more easily investigate the effect of scale in isolation from other confounding factors.

- **Qwen3** (Qwen Team, 2025): A suite of state-of-the-art decoder-only transformers released by Alibaba Cloud in 2025 and trained on a large multilingual corpus. We use five sizes—0.6B, 1.7B, 4B, 8B, and 14B—each in two variants: a *base* model (pre-training only) and an *instruct* model (base model further trained with supervised fine-tuning and reinforcement learning from human feedback). This 5×2 design allows us to investigate the effect of post-training independently of model scale.

- **Sentence Gestalt (SG) model** (Lopopolo & Rabovsky, 2024): An LSTM-based model, which, unlike the next-word prediction models above, is trained on a qualitatively different task: mapping sentences incrementally to their thematic event structure, represented as a set of role–filler pairs (e.g., ⟨AGENT, *boy*⟩, ⟨ACTION, *open*⟩, ⟨PATIENT, *door*⟩) encompassing 26 argument types. The model was trained on approximately 8 million tokens of annotated English narrative text, with role fillers represented using FastText embeddings. The SG model consists of a 600-dimensional word embedding layer feeding a single-layer, 1,200-unit LSTM (the sentence gestalt layer), which maintains a running implicit representation, useful for the role–filler task, that is updated word-by-word. We used Lopopolo & Rabovsky's pre-trained model for our investigations.

3.3. Model Predictors

From these models, we computed a wide range of metrics which we then used as predictors of human N400 responses to the same sentences. Those metrics are as follows:

- **Surprisal**: The surprisal of the critical word under the model's final-layer output distribution, $-\log_2 P(w_t \mid w_1, \ldots, w_{t-1})$. This is the standard computational-level predictor in the psycholinguistic literature, reflecting how unexpected a word is given its full preceding context. It is available for all of the transformer models we compared. Among other papers, Michaelov et al. (2024) argue that this value provides the best computational account of the N400.

- **Layer-wise surprisal**: Surprisal extracted from the transformer layer that best predicts N400 amplitude, selected via leave-one-dataset-out cross-validation: for each held-out dataset, the optimal layer was identified by maximizing the absolute correlation

$|r(\text{surprisal}, \text{N400})|$ across training datasets; test surprisal on the held-out dataset was then calculated based on the predictions from that layer. Per-layer distributions were obtained using pre-trained tuned-lens linear probes (Belrose et al., 2023), which are linear layers trained at each layer to approximate the final next-token distribution. This allows us to translate each intermediate hidden state into a vocabulary distribution. This layer-wise surprisal is available only for models with published tuned-lens probes (GPT-2 and Pythia families), so we do not report these values for any of the Qwen3 models. We decided to include this predictor as intermediate layers may contain more abstract semantic information than the final layers of a model.

- **Shallow surprisal**: Based on pyscholinguitsic accounts of the differences between the N400 and P600 ERPs as "shallow" versus "deep" processing, Li & Futrell (2024) propose an N400 predictor defined as the KL divergence $D_{\mathrm{KL}}(p_\lambda \| p_0)$ between a distribution p_λ and the model's base distribution p_0. To construct p_λ, we reweight p_0 as follows: $p_\lambda(w) \propto p_0(w)\ \exp[-\lambda(\mathrm{Lev}(w, x) + \gamma \cos(e_w, e_x))]$. This upweights predictions toward vocabulary items that are orthographically similar (Levenshtein distance) and semantically similar (cosine distance between token embeddings) to the actual next token. We used $\lambda = 1.0$ and $\gamma = 8$, following the original paper. The value $D_{\mathrm{KL}}(p_\lambda \| p_0)$ then intuitively captures the extent to which the model's predictions were "almost right"; higher values here mean that the model's top predictions were farther away in orthographic and embedding space from the actual next token. We can compute this value for all transformer models.

- **Language-network update**: Inspired by AlKhamissi et al. (2025) and the "functional localizer" approach in cognitive neuroscience (e.g. Fedorenko et al., 2011), we decided to include a metric computed from units in the transformer models which are critically important for language processing. To identify language-selective units within each model, we sampled 100 sentences and 100 nonword sequences from the Fedorenko localizer corpus, extracted mean-pooled hidden-state activations across all layers for each stimulus, and selected the top 1% of units (across the flattened layer $\times$ embedding-dimension space) by their sentences $>$ nonwords activations. These units were shown by AlKhamissi et al. (2025) to be critically important for language production, as well as better predictors of fMRI-defined language network activity in humans. For each of our experimental stimuli, we extracted activations of these language-network units at the token immediately preceding the critical word and at then the final subword token of the critical word, considering both as vectors, and computed the mean absolute difference between these two vectors. We then Z-scored these values within each model before analysis to remove scale differences across architectures of different depths and widths. Intuitively, this metric corresponds to how much the model's internal states, defined over the language network units, has to change when a new word is encountered. We can compute this value for all transformer models. We chose the mean absolute difference value instead of the more-standard cosine distance for direct comparison with the SG model update metric, discussed next.

- **Sentence Gestalt update**: Lopopolo & Rabovsky (2024) propose a different representation-change metric, computed from their SG model described above, as a predictor of the N400. At each word position i, the model updates its gestalt state $\mathbf{sg}_i \in \mathbb{R}^{1200}$; the update value is the mean absolute change in that state relative to the preceding position:

$$\text{update}_i = \frac{1}{H} \sum_{h=1}^{H} \left| \text{sg}_{i,h} - \text{sg}_{i-1,h} \right|, \quad H = 1{,}200.$$

  This value measures how much the model's representation of the sentence must change when encountering a new word. This is analogous to the Language-network update metric described for transformers above, but computed from the internals of a different model architecture trained with a different objective. We also Z-score these values before further analysis. This value, by definition, can only be computed for Lopopolo & Rabovsky's SG model.

3.4. Statistical Analyses

We assess the predictive value of these different metrics by fitting LME models using each metric as a predictor of the N400 amplitudes, then comparing this model to a null model. To be more precise, we first fit a full model for each metric:

$$\textbf{N400} \sim \textbf{\textit{metric}} + \{\text{controls}\} + (1|\text{subject}) + (1|\text{stimulus})$$

And a null model:

$$\textbf{N400} \sim \{\text{controls}\} + (1|\text{subject}) + (1|\text{stimulus})$$

With the controls being word frequency, word length, and the position of the word in the sentence, which are all known to impact N400 amplitudes but are not our manipulations of interest (Michaelov et al., 2024). Note that neither Lopopolo & Rabovsky (2024) nor Li & Futrell (2024) include these controls in their analyses.

Both models were estimated by maximum likelihood (`REML = FALSE`) to permit a likelihood ratio test (LRT) using `lme4` and `lmerTest`. The test statistic $\chi^2(1) =$

$2\,(\log L_{\text{full}} - \log L_{\text{null}})$ follows a χ^2 distribution with one degree of freedom under the null hypothesis that the predictor explains no additional variance. We additionally report $\Delta\text{AIC} = \text{AIC}_{\text{null}} - \text{AIC}_{\text{full}}$ as a sample-size-penalised effect-size measure; positive values indicate that the full model is a better predictor of the N400 amplitude across datasets, participants, and trials than the null model. We can then compare these ΔAIC values across different metrics to assess their relative predictive values.

4. Results

4.1. Comparing Predictors

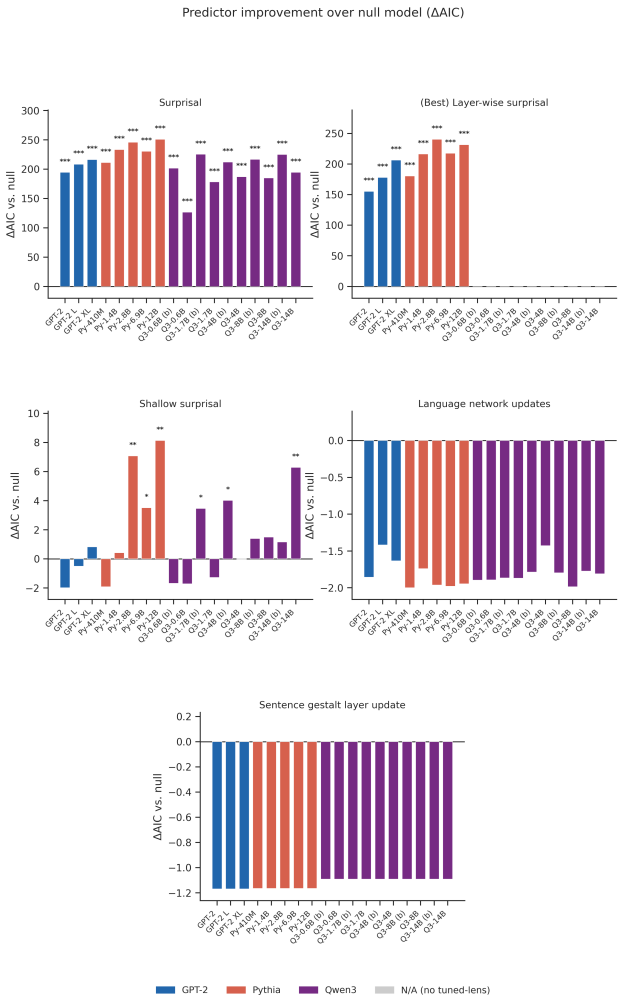

Figure 1. ΔAIC values for full-versus-null models including metrics computed from all language models tested across the full dataset. Significance stars reflect the likelihood ratio test p-value (*** < 0.001, ** < 0.01, * < 0.05). Qwen models with (b) included in their name reflect base models, the unmarked Qwen models reflect instruct versions.

Figure 1 shows the ΔAIC values computed for all full-vs-null model comparisons across all language models on the whole dataset.

Contrary to our initial hypotheses, the transformer-based surprisal metrics, across all applicable langauge models, computed from both the final layer and the leave-one-dataset-out cross-validated internal layers, have large ΔAIC values ($\Delta\text{AIC} > 150$, $p < 0.001$), showing very strong support for the full models over the null model. The shallow surprisal metric has lower ΔAIC values, and is only significant for some lanugage models ($10 > \Delta\text{AIC} > 0$, $p < 0.05$), indicating significant but weaker support for the shallow-surprisal metric over the null model. These results support the conclusions of both Michaelov et al. (2024) and Li & Futrell (2024).

The two state-update metrics (both computed from the transformer language-networks and the SG model), however, do not reflect any predictivity improvement over the null model ($\Delta\text{AIC} < 0$). (Note that differences in tokenization between the GPT2/Pythia models (which use the same tokenizer) and the Qwen models leads to a different null model baseline (due to differences in the word-length control), which is why the SG update plot shows different values for the Qwen models). We will return to reasons why these metrics may not have fared so well on our benchmark in the discussion section below.

4.2. Layer-wise Surprisal

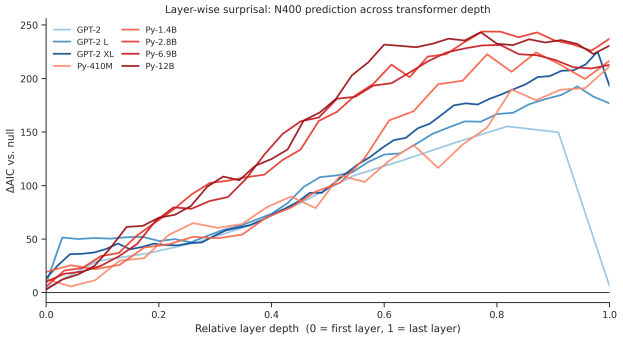

Figure 2. ΔAIC values for full-versus-null models including metrics computed from transformer language model surprisal, across model depth. Colors reflect model families, hues reflect model size.

Now that we know that layer-wise surprisal, computed from the cross-validated best layer, is a strong predictor of N400 amplitude, what layers tend to be selected as this best layer?

Figure 2 shows how ΔAIC changes as a function of the layer we compute the surprisal from, across model sizes within

the GPT2 and Pythia families. ΔAIC values continue to climb across model depth for most language models, while for the largest (and best-predicting models) this increase begins to plateau after around 60% of a model's depth.

4.3. Scaling, Model Family

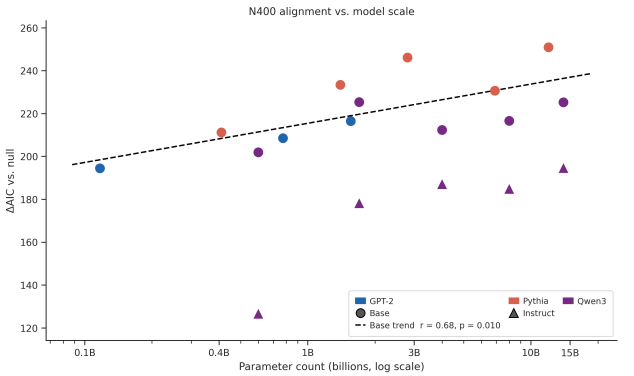

Figure 3. ΔAIC values for last-layer surprisal across model sizes and families. Colors reflect model families, shapes reflect training type (base=pre-trained only).

As advances in hardware and model infrastructure have made ever-larger LLMs possible in recent years, much attention has been directed towards studying how model scale affects a wide range of downstream behaviors (Kaplan et. al, 2020; Antonello et al., 2023). Here, we present a novel result of examining these "scaling laws" in relation to the N400.

Figure 3 shows the main results of this investigation: using last-layer surprisal as our metric of interest, we can see that the ΔAIC value steadily climbs in relation to the model parameter count (log scale). There is a significant and strong correlation between this parameter count and ΔAIC (Pearson r = 0.68, $p < 0.05$), at least for the pre-trained-only base models (the circles in Figure 3).

We see a notably distinct pattern when we compare these base models to the Qwen3 Instruct models (the triangles in Figure 3), which have been post-trained with supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) for human-preference-alignment and instruction following. While it appears their N400 predictivity also increases with scale, the baseline ΔAIC values are significantly lower than their size-matched, base-model Qwen3 counterparts. Note that this appears to be a result specifically of this post-training, rather than the pre-training of Qwen3 as opposed to other model families, as the base Qwen3 models fall along the same trend line as the other base models from other model families. We will return to the implications of this result for human-model alignment in the discussion section below.

5. Discussion

5.1. Comparing competing computational accounts

The central aim of this work was to provide a standardized benchmark for evaluating competing computational accounts of the N400 ERP, one of psycholinguistics' most well-studied neural measures of human semantic processing. To accomplish this, we compiled and standardized the results from three separate laboratory experiments, totaling nearly 100 individual participants and 46,000 individual trials, into a novel dataset we call EEGym. We then used this dataset for initial investigations to compare various computational methods for modeling the N400 response with LLMs across a wide range of candidate models.

Our results offer several points of clarification regarding the ongoing debate in the literature (see Figure 1). Most broadly, we find that transformer-based surprisal–the negative log probability of a word given its preceding context–is the strongest predictor of N400 amplitude across all three datasets in EEGym, with large and highly significant ΔAIC values ($\Delta AIC > 150$, $p < 0.001$) for all transformer models tested (with the exception of the smallest Qwen3 Instruct model, which had a marginally lower ΔAIC). We find that surprisals computed from both the final layer and mid-to-late layers in transformer models are strong predictors. This result is consistent with and extends the conclusions of Michaelov et al. (2024). Further, our benchmark provides a more stringent test of that claim: whereas prior work established the surprisal-N400 link within a single experimental paradigm, here we replicate it across both naturalistic continuous reading (Frank et al., 2015) and controlled anomalous-sentence paradigms (Ryskin et al., 2021; Michaelov et al., 2024), suggesting the relationship is reasonably robust to variation in experimental design.

Contrary to our initial hypothesis, we did not find evidence that more cognitively-inspired metrics outperform the more-standard surprisal. The shallow surprisal metric proposed by Li & Futrell (2024), while statistically significant for some models, showed considerably weaker predictivity than standard surprisal (ΔAIC 2–8 versus > 150). One possible reason for this difference is that Li & Futrell's original analyses did not include the word frequency, length, and position controls, and that the shallow surprisal metric captures variance that is partially explained by these lower-level stimulus-properties. We therefore interpret these results as a qualified replication of Li & Futrell (2024): shallow surprisal appears to contribute something beyond what these controls alone capture, but less than what standard surprisal contributes. A further reason for the difference between our results and those in Li & Futrell is the dataset composi-

tion: they compare responses only in controlled, anomalous-sentence paradigms, while here we are pooling across this kind of design and naturalistic reading. Finally, their model contains one free parameter, γ, which they fit to their experimental stimuli. To maximize predictive power for our main analyses, we did not hold any stimuli out for this parameter-fitting, and instead used the value for γ they find best fits to the Ryskin data in their analyses.

We also examined cognitively-inspired metrics based on representational-state-updating: the Sentence Gestalt model update proposed by Lopopolo & Rabovsky (2024) and an analogous metric computed from the transformers' "language-network" units, identified in a manner comparable with human fMRI investigations of language (AlKhamissi et al., 2025). Neither of these metrics added any predictive value above that of our control variables (ΔAIC < 0). We offer three tentative explanations. First, as with shallow surprisal, the absence of frequency, length, and position controls in the original paper by Lopopolo & Rabovsky (2024) may play a role in the discrepancy. Second, again like the shallow surprisal discussion above, the SG model update was proposed as an account of naturalistic reading data, and here we are pooling across naturalistic and controlled paradigms. Third, it is possible that the mean absolute difference measure, while a reasonable operationalization of representational change, does not optimally capture the computations underlying the N400. Future work might explore alternative distance metrics or representations. We are cautious, however, about drawing strong theoretical conclusions from null results, and we acknowledge that the benchmark we propose here, though larger than those used in prior individual studies, may still lack the statistical power to detect small but genuine effects.

5.2. Investigations of model depth and scale

Moving into more specific investigations of the relationship between model representations and the N400, our layer-wise surprisal analyses reveal that mid-to-late layers (at around 60% of a model's depth) tend to be the strongest predictors of N400 amplitude, with predictivity plateauing or declining slightly in the final layers for the largest models (see Figure 2). This is consistent with the view that intermediate representations may encode more abstract semantic information than the final layers, and consistent with these abstract representations being a better computational account of the N400, as discussed in the introduction. This result is suggestive, though we note that the layer selection procedure used here (leave-one-dataset-out cross-validation) introduces its own sources of variance, and the optimal layer varied somewhat across datasets. More systematic investigation of the relationship between layer depth, N400 predictivity, and experimental paradigm would help clarify this pattern.

As last of our main results, we find that N400 predictivity is strongly correlated with model size (see Figure 3). This scaling relationship holds across the GPT2, Pythia, and Qwen3 model families, suggesting it is a relatively general property of transformer models rather than an artifact of a particular architecture or training corpus. Importantly, however, the scaling trend appears to be modulated by training type: base (pre-trained-only) models consistently outperform their instruction-tuned counterparts at equivalent parameter counts, at least for the Qwen3 family. This finding is, to our knowledge, novel in the context of N400 prediction. One plausible interpretation is that models that produce more human-preferred outputs do not necessarily develop more human-like online language processing, at least as is indexed by the N400. This dissociation between preference alignment and neural predictivity is theoretically important and deserves further investigation, including examination of what specific aspects of post-training are responsible for the reduction in N400 predictivity.

5.3. Potential cognitive interpretations

Finally, we offer one exploratory, tentative note regarding the differences between these results and the more well-studied comparisons between LLM representations and human fMRI responses. First, AlKhamissi et al. (2025) find that representations from transformer "languge network" units are more predictive of fMRI-defined human language network activity than an equally-sized random subset of units; second, Tuckute et al. (2024) find that fMRI encoding performance increases dramatically from the embedding layer to early layers in GPT2XL then plateaus at about 25-50% of the model's depth before declining; third, Aw et al. (2023) find that instruction-tuned models are slightly *more* predictive of fMRI activity than base models. While LLM-to-fMRI encoding models are a wholly different measure from the ΔAIC metric we use here, one tentative interpretation of the discrepancy between these sets of results and ours is that the N400 and fMRI-defined language network activity are measures of two *different* neural computations during online language processing. This would be an important result in the cognitive neuroscience of language, and deserves a much more rigorous investigation than what we can show here. However, we hope to follow up on this in future work.

5.4. Limitations

While we constructed a dataset much larger than is typically used for LLM-EEG comparison studies, the generalizability of our results is of course limited by the size and generalizability of our benchmark. Open-science practices have become widely adopted relatively recently in cognitive neuroscience, and much of the hallmark work on the N400 was done prior to this development, making dataset aggregation

somewhat difficult. Many of these important studies only reported condition averages rather than the specific trial-wise data we need to do a rigorous comparison of models and theories. Additionally, we were somewhat limited in our computational expressivity by not re-fitting the free parameters in the shallow surprisal model (discussed above) or training our own tuned-lens linear models for the Qwen3 family. Finally, as briefly mentioned above, a more direct comparison with the literature on fMRI predictivity would be afforded by taking an encoding-model approach, where we use LLM representations to predict the EEG response. These would be important computational next-steps for an even more generalizable comparison with previous results.

6. Conclusion

We present an initial version of EEGym: an effort to collect and standardize EEG reading datasets for more rigorous model-brain comparisons in NLP and cognitive neuroscience. We additionally conduct some initial investigations on this new benchmark; taken together, the our suggest that transformer surprisal remains the most robust computational predictor of the N400 in the current literature, despite growing theoretical interest in more cognitively-inspired alternatives. This does not preclude the possibility that future metrics, will prove superior, perhaps drawing on richer representations or trained on more cognitively relevant objectives. The present benchmark is intended precisely to facilitate such future comparisons. We release EEGym as an open resource and hope it will serve as a common testbed for future work at the intersection of NLP and cognitive neuroscience.

7. Data and Code Availability

Cleaned datasets, analysis code, and interactive figures can be found at: https://samhutch511.github.io/nlp-project/. The full GitHub repo can be found at: https://github.com/jen-baek/nlp-group-project.

References

AlKhamissi, B., Tuckute, G., Bosselut, A., and Schrimpf, M. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Association for Computational Linguistics*, 2025.

Antonello, R. J., Vaidya, A. R., and Huth, A. G. Scaling laws for language encoding models in fMRI. In *Advances in Neural Information Processing Systems*, volume 36, pp. 21895–21907, 2023.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., and Crocker, M. W. Retrieval (N400) and integration (P600) in expectation-based comprehension. *PLOS ONE*, 16(9): e0257430, 2021.

Aw, K. L. and Toneva, M. Instruction-tuning aligns LLMs to the human brain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Caucheteux, C., Gramfort, A., and King, J. R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441, 2023.

De Varda, A. G., Marelli, M., and Amenta, S. Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5): 5190–5213, 2024.

Eddine, S. N., Brothers, T., and Kuperberg, G. The N400 in silico: A review of computational models. *Psychology of Learning and Motivation*, 76:123–206, 2022.

Fedorenko, E., Behr, M. K., and Kanwisher, N. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433, 2011. doi: 10.1073/pnas.1112937108.

Frank, S. L., Otten, L. J., Galli, G., and Vigliocco, G. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11, 2015.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Krieger, B., Brouwer, H., Aurnhammer, C., and Crocker, M. W. On the limits of LLM surprisal as a functional explanation of the N400 and P600. *Brain Research*, 1865, 2025.

Kutas, M. and Federmeier, K. D. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, 62(1):621–647, 2011.

Kutas, M. and Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163, 1984.

Li, J. and Futrell, R. Decomposition of surprisal: Unified computational model of ERP components in language processing. *arXiv preprint arXiv:2409.06803*, 2024.

Lopopolo, A. and Rabovsky, M. Tracking lexical and semantic prediction error underlying the N400 using artificial neural network models of sentence processing. *Neurobiology of Language*, 5(1):136–166, 2024.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., and Coulson, S. Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 5(1):107–135, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

Rugg, M. D. The effects of semantic priming and word repetition on event-related potentials. *Psychophysiology*, 22(6):642–647, 1985.

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., and Gibson, E. An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, 158:107855, 2021.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., and Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), 2021.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., and Fedorenko, E. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.